BMC Bioinformatics

RESEARCH ARTICLE Open Access

Content-based microarray search using differential expression profiles

Jesse M Engreitz[1], Alexander A Morgan[2], Joel T Dudley[2,3], Rong Chen[3], Rahul Thathoo[4], Russ B Altman[1,5], Atul J Butte[2,3,6*]

Abstract

Background: With the expansion of public repositories such as the Gene Expression Omnibus (GEO), we are rapidly cataloging cellular transcriptional responses to diverse experimental conditions. Methods that query these repositories based on gene expression content, rather than textual annotations, may enable more effective experiment retrieval as well as the discovery of novel associations between drugs, diseases, and other perturbations.

Results: We develop methods to retrieve gene expression experiments that differentially express the same transcriptional programs as a query experiment. Avoiding thresholds, we generate differential expression profiles that include a score for each gene measured in an experiment. We use existing and novel dimension reduction and correlation measures to rank relevant experiments in an entirely data-driven manner, allowing emergent features of the data to drive the results. A combination of matrix decomposition and *p*-weighted Pearson correlation proves the most suitable for comparing differential expression profiles. We apply this method to index all GEO DataSets, and demonstrate the utility of our approach by identifying pathways and conditions relevant to transcription factors Nanog and FoxO3.

Conclusions: Content-based gene expression search generates relevant hypotheses for biological inquiry. Experiments across platforms, tissue types, and protocols inform the analysis of new datasets.

Background

With the development of the DNA microarray and other technologies that probe gene expression on an "omic" scale, we are now able to discover associations between biological conditions based on their molecular underpinnings. Seminal work by Golub et al. [1] classified leukemia samples by their global gene expression profiles, demonstrating that transcriptomic signatures can aid in functional prediction and improve our molecular understanding of disease. Hughes et al. [2] predicted the effects of novel gene deletions and chemical treatments by profiling yeast mutants and comparing new arrays to this reference. More recent studies examined cellular transcriptional response to drug treatment [3,4] and disease [5,6] in order to identify novel relationships between apparently unrelated conditions and compounds. This work not only demonstrated the utility of expression-based discovery, but also suggested that functional studies about drugs and diseases can utilize data from different platforms and cell types. This general approach to hypothesis generation - namely, finding associations between diverse conditions based on gene expression - has great potential to further biological and biomedical research if implemented on a large scale.

Here we develop methods for content-based gene expression search using an entire experiment as a query. That is, given an input experiment comparing case to control, we aim to identify other experiments that show similar patterns of differential expression. This concept is exemplified by the Connectivity Map [3], which searches for relationships between treatment-control comparisons for small molecules. While the Connectivity Map focused on drug treatment and disease, a similar approach across a sufficiently large data source would allow for the identification of associations between gene knockdowns, diseases, drugs, and myriad

* Correspondence: abutte@stanford.edu
[2]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA, USA
Full list of author information is available at the end of the article

other perturbations and phenotypes. Public repositories provide a wealth of data amenable to this task. The largest of these repositories, the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) [7], now contains over 400,000 individual samples from more than 17,000 experiments detailing the molecular characteristics of diverse cell types, diseases, and drug treatments. The European Bioinformatics Institute (EBI) ArrayExpress Repository [8] and Stanford Microarray Database [9] host additional data. While GEO supports searches of its content based on free-text and controlled-vocabulary annotations, there is increasing interest in methods for querying microarray databases based on the molecular measurements themselves [10-14]. The power of this approach would grow with the size of the repository.

Current methods for content-based search typically involve a two-step process: they identify a gene set of interest and then search for experiments in which this gene set is important. Several groups have introduced methods for identifying experiments that co-express [15] or differentially express [11] a given gene set. Recently, EBI implemented the Gene Expression Atlas, which provides this latter functionality over their curated array archive [13]. These methods, however, require that both the query and target experiments differentially express genes above some hard threshold, and thus may miss more subtle or noisy relationships [16]. Other approaches, typified by Gene Set Enrichment Analysis (GSEA) [16], partially bypass this requirement by comparing a subset of genes to ranked profiles, using a hard threshold for the query experiment and a soft threshold for the queried experiments [3,4].

While previous approaches require designating a group of differentially expressed genes, we explore the possibility of using as a query a differential expression (DE) profile, consisting of a complete list of features and associated expression scores. By examining all genes shared between query and queried experiments, we aim to identify experimental conditions and perturbations that exhibit similar transcriptional responses. A successful strategy in this effort should reconcile differences between species, platform types, and normalization methods, as well as overcome the confounding effects of noise and technical replicability. To achieve this, we consider combinations of methods for three tasks: data representation, dimension reduction, and search algorithm.

First, we consider the problem of data representation. Typical microarray analysis methods represent differential expression as a fold-change, comparing the expression in one set of samples to that in another [17]. However, because public expression databases consist of a broad range of data types and experimental modalities, rank-based representations are often employed to account for the disparities in the distributions of observed data [4,10]. Here we compare both parametric and nonparametric data representations to determine the best approach for comparing DE profiles. We also consider an alternate representation of gene expression data, and construct DE profiles based on the p-value of differential expression.

A second challenge is that gene expression profiles from high-throughput technologies consist of up to tens of thousands of measurements per sample. In addition to the computational complexity involved in handling these large datasets, high dimensionality often confounds data mining techniques [16,18]. In particular, high-dimensional, multimodal data lends itself to overfitting and reduced performance [19]. Many solutions to this problem have been proposed, of which dimensionality reduction is the foremost. Matrix decomposition [20,21], feature selection [22], and module or gene-set based approaches [16,23] attempt to capture the most relevant data while removing redundant or noisy features [18].

Given an appropriate data representation for differential expression, the final challenge is how best to calculate the similarity between two experiments. While Fujibuchi et al. use Spearman rank correlation to compare individual microarrays [10], it is not clear whether a similar approach is appropriate for DE profiles. Several recent studies use a modified Pearson correlation measure on rank-normalized profiles [4,5,24]. Other work suggests that weighting expression values by each gene's variance may improve classification and analysis [25,26].

To begin to address these challenges, we test several search schema representing combinations of data representation, dimension reduction, and correlation measures in a curated collection of 32 disease-related GEO experiments. We create DE profiles to represent the changes in transcription between normal and disease samples (Figure 1A), and evaluate the performance of our schema in retrieving experiments that measure the same disease as a query experiment (Figure 1B). We find that a projection method for dimension reduction performs as well or better than search in gene-space, and introduce an intuitive p-value weighted correlation coefficient that performs the best in our test compendium. Using the most successful parameters, we exhaustively index GEO DataSets (GDS) totaling 31,453 arrays and 2,089 experiments. We demonstrate the utility of our method by querying our database of DE profiles with several experiments examining transcription factor knockdown in embryonic and neural stem cells. This work demonstrates the feasibility of content-based microarray search for the large-scale discovery of functional links between gene expression experiments.

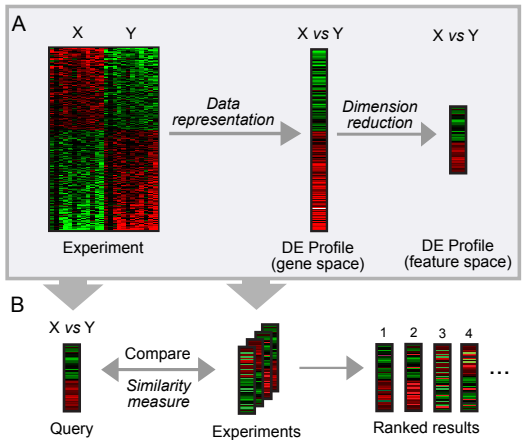

Figure 1 Schematic diagram of our approach. (A) Creation of differential expression (DE) profiles. An experiment comparing condition X with condition Y is condensed to a single DE profile in gene space. Dimension reduction is applied to create a DE profile in a reduced feature space. (B) Searching a library of DE profiles. The query profile is compared to the DE profiles of all other experiments in our disease compendium (or in GEO) using a similarity measure. Results are ranked by their similarity to the query profile. Italics indicate variable steps in our pipeline.

Results

Evaluation of data representation and similarity measures

To develop a differential-expression search utility for GEO, we first evaluated various data processing pipelines in a compendium of 32 microarray experiments comparing normal to diseased tissue. This collection included three diseases with differing genetic origin: Duchenne muscular dystrophy, Huntington's disease, and breast cancer. The studies originated from different laboratories and measured primary human disease samples as well as animal disease models. Although these experiments represented various combinations of species, platform, and normalization techniques, they clustered primarily by disease and tissue (Figure 2). To search this collection of experiments based on differential expression (DE), we created a DE profile for each experiment (32 total), consisting of a list of features (e.g., genes) each with an associated score (e.g., fold-change). We permuted various processing and ranking techniques to search for the combination of parameters that was best able to identify other experiments of the same disease given a query experiment. We evaluated the sensitivity and specificity of these processing pipelines with leave-one-out cross-validation: we used each experiment to query the remaining 31 experiments with the goal of identifying other experiments that measure the same disease.

First we compared the effects of data representation on our ability to retrieve relevant experiments. Using both Pearson and Spearman correlation, we found that representing differential expression as a log fold-change nominally outperformed rank- and p-value-based representations (see Additional file 1). For subsequent tests we focused on the fold-change representation. Next we evaluated all combinations of four dimension reduction methods and six similarity measures (Figure 3). For dimension reduction, projection onto features identified by independent component analysis (ICA, see Methods) [27] outperformed module-based representations. While none of the dimension reduction methods made convincing improvements over the gene-level analysis, the ICA projection method did not result in a loss of information, successfully recapitulating performance in gene-space using a significantly reduced number of features. For similarity measures, unweighted and p-value weighted Pearson correlations nominally outperformed Spearman correlation for analysis in gene- and ICA feature-space, resulting in the highest overall areas under the receiver operating curve. P-weighted Spearman correlation performed the worst for all dimension reduction methods.

The receiver operating curves for the best-performing search methods indicated that, on average, about 50% of the true positives could be recovered with greater than 90% specificity. This high specificity is important for search because typically the first few results, rather than a complete list, are examined. To evaluate the performance of our search over the top results in each search, we calculated the "precision at 4" for each of the 32 experiments, permuting labels to create a null model (see Additional file 2). The average precision for Duchenne muscular dystrophy and Huntington's disease exceeded the random model at a 95% confidence interval for 13/15 and 8/8 experiments, respectively. The "precision at 4" for breast cancer, a genetically complex disease, was also high, but it significantly surpassed the random model in only 4/9 experiments.

Constructing a network of GEO differential expression experiments

Our comparisons of data processing methods and similarity measures suggested that the p-weighted Pearson correlation in gene- or ICA-space is most effective at retrieving biologically relevant DE profiles. Because the ICA-based method reduced the number of features by a factor of 50, we used this approach to systematically index GEO DataSets. We created a total of 9,415 DE profiles, one for each combination of NCBI-curated experimental conditions within a dataset. For example, if a dataset had "1 hr", "2 hr", and "4 hr" groups, we generated a comparison for each of "1 hr vs 2 hr", "2 hr vs 4 hr", and "1 hr vs 4 hr." We excluded 364 comparisons that failed to successfully map to human genes through Homologene; these experiments measured

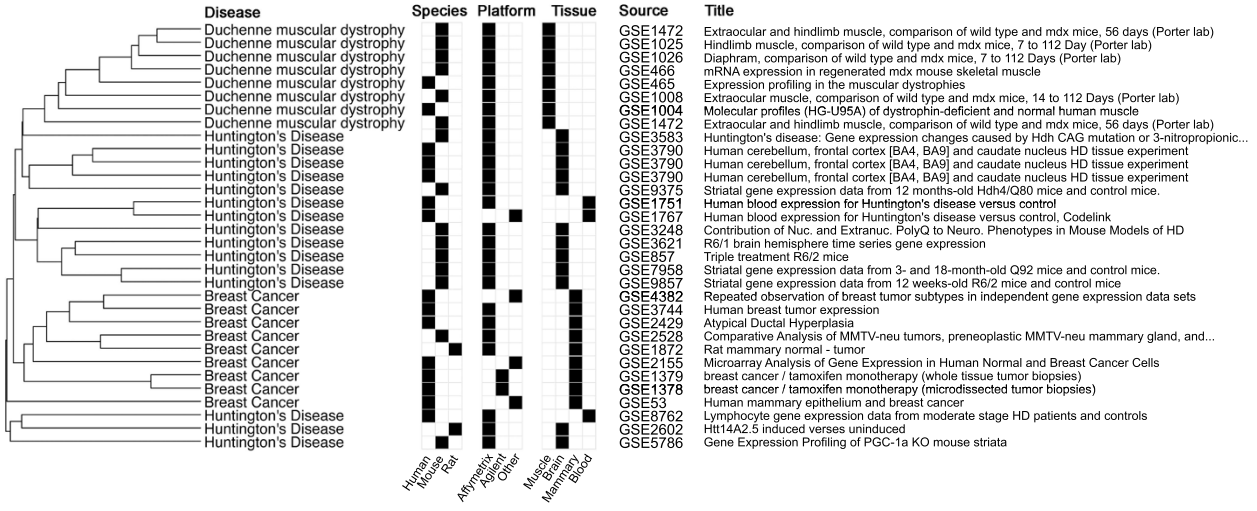

Figure 2 Disease compendium. Our collection of 32 disease-related experiments represents several combinations of species, platforms, and tissues. Differential expression profiles based on log fold-change were generated for each experiment, then mapped to human genes through Homologene. We compared DE profiles using Pearson correlation and applied hierarchical clustering to find that the profiles cluster primarily by disease and tissue. One GEO Series appears more than once: GSE3790 provides three profiles that cluster together, comparing normal to diseased tissue in cerebellum, frontal cortex, and caudate nucleus.

primarily bacterial and plant species. To visualize this set of profiles, we calculated pairwise similarities using p-weighted Pearson correlation and created a network of differential expression experiments (Figure 4A, Additional files 3, 4). Random comparisons were used to build a null distribution of similarity scores. With a strict cutoff ($q < 0.001$), highly-connected subnetworks consisting of multiple profiles from the same dataset emerged. Clusters of profiles from multiple experiments also were apparent, linking datasets that examined related biological processes and perturbations. Figure 4B shows a multi-experiment cluster examining gonad development in mouse, consisting of differential expression profiles from GDS2098, GDS2203, and GDS2719. Each profile compares gonad tissue at two developmental stages, between 10 and 18 days post coitum. The highly significant associations between the testis (GDS2098) and ovary (GDS2203) reflect known molecular similarities between male and female gonad development, especially before gestation day 10.5 [28,29]. Profiles comparing later stages in development are not linked between the sexes (e.g., starred profile in Figure 4B).

Application to Nanog knockdown in embryonic stem cells

This search method allowed us to simultaneously investigate a wide range of perturbations, conditions, and comparisons using the hypothesis that experiments that differentially express similar genes and pathways would also share functional phenotypic relationships. To assess the utility of this approach, we used data from GDS1824 to investigate the effects of Nanog knockdown in embryonic stem cells (ESCs) [30]. We created a DE profile comparing Nanog knockdown to control in mouse ESCs, and queried all GEO DataSets to identify other experiments that have similar differential expression patterns. Because the transcription factor Nanog is required for the maintenance of pluripotency in ESCs [31,32], we hypothesized that this search would find profiles comparing embryonic stem cells to differentiated cells. Indeed, ten of the top fifteen matching profiles consisted of experiments comparing less differentiated to more differentiated mouse embryoid bodies of various genetic lineages (Figure 5). In matching experiments, differentiation was induced by removal of LIF (leukemia inhibitory factor) [33], a cytokine necessary to maintain the undifferentiated state of ESCs [34]. The Nanog knockdown search also identified comparisons from GDS1823, also from Loh et al. [30], where ESC differentiation was induced by drug treatment with retinoic acid (RA) or hexa-methylene-bis-acetamide (HMBA).

In addition to mouse ESC datasets, this search produced interesting comparisons with different experimental systems. Result 14 supports a similarity between Nanog knockdown and the comparison of non-small cell lung carcinoma (NSCLC) to small cell lung cancer (SCLC). SCLC, the more aggressive disease, has been linked with expression of stem cell factor [35] and the Hedgehog signaling pathway [36]. These relationships suggest that, in a broad sense, SCLC compared to

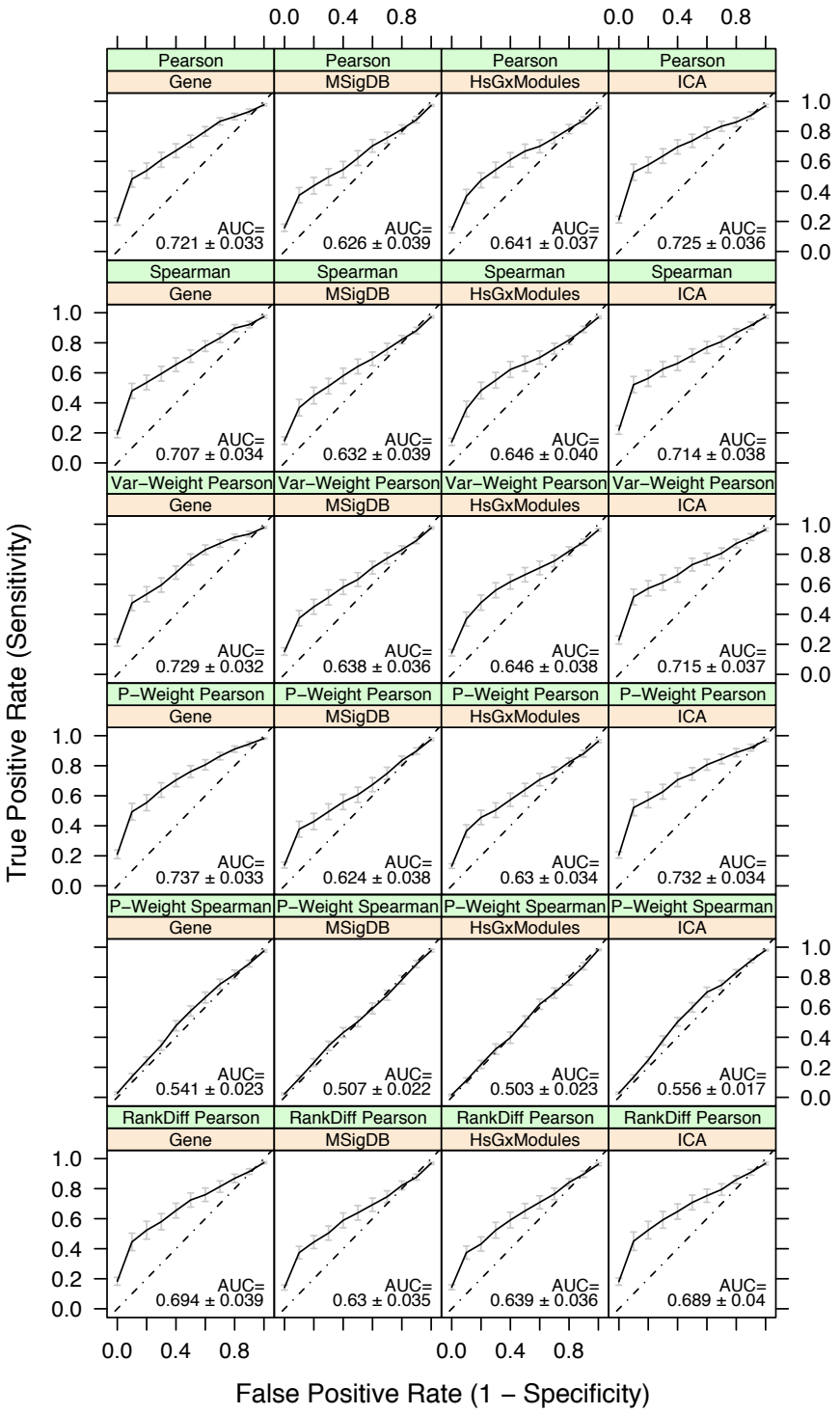

Figure 3 Evaluation of dimension reduction methods and similarity measures. Comparison of four dimension reduction methods and six similarity measures using leave-one-out cross-validation in our disease compendium. Bars and AUC estimates indicate standard errors for curves averaged over all cross-validation trials. The three similarity measures based on Pearson correlation outperform the rank-based approaches, with the *p*-weighted Pearson correlation proving the best at identifying other experiments of the same disease. The ICA projection method for dimension reduction outperforms the module-based approaches, and performs comparably to gene-level analysis. HsGxModules = Human Gene Expression Modules (see Methods).

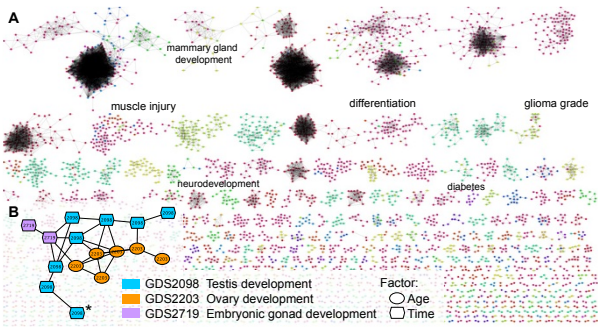

Figure 4 Network of GEO differential expression profiles. (A) We calculated *p*-weighted correlations between 9,415 differential expression profiles from GEO and connected highly similar profiles ($q < 0.001$). Nodes are colored according to experimental variable (e.g., time). Dense clusters tend to represent multiple profiles from the same experiment. We identified multi-experiment clusters corresponding to processes including muscle injury, mammary gland development, and glioma grade. For a high resolution figure, see Additional files 3 and 4. (B) Close-up of a multi-experiment cluster. DE profile nodes are re-colored to correspond to the GEO DataSet from which they originate, and node shape represents experimental variables. *Compares gestation day 14 to gestation day 16.

NSCLC may have a more stem-like transcriptional program.

Our method also identifies the genes that drive the correlation between two profiles. These genes have the most significant coordinated changes in the two experiments. When we examined the genes driving the correlation for Result 14, we found cytokeratin KRT18 overexpressed in both Nanog knockdown compared to control and in NSCLC compared to SCLC (Figure 6B), fitting previous examinations of KRT18 by immunohistochemistry [37,38]. On the other end of the spectrum, we found the relatively uncharacterized gene FXYD6, a regulator of N, K-ATPase [39]. FXYD6 is down-regulated during Nanog knockdown (see Additional file 5) and up-regulated in several SCLC cell lines (see Additional file 6). This suggests that FXYD6 plays a role in the transcriptional programs in common between embryonic stem cells and SCLC.

Query: GDS1824 Transcription factors Nanog and Oct4 knockdown effect on embryonic stem cells ESC (control) *vs.* ESC (Nanog knockdown)

Rank	GEO	Title	Subset 1	*vs.*	Subset 2	Type	Score	*q*-value
1	GDS2668	Embryonic J1 stem cell differentiation in vitro (MG-430A)	24 h	*vs.*	4 d	time	0.596	0.0054
2	GDS2668	Embryonic J1 stem cell differentiation in vitro (MG-430A)	36 h	*vs.*	4 d	time	0.565	0.0073
3	GDS2668	Embryonic J1 stem cell differentiation in vitro (MG-430A)	18 h	*vs.*	4 d	time	0.558	0.0081
4	GDS2668	Embryonic J1 stem cell differentiation in vitro (MG-430A)	24 h	*vs.*	7 d	time	0.551	0.0087
5	GDS1823	Embryonic stem cell differentiation induced by various chemicals: time course	DMSO	*vs.*	HMBA	agent	0.539	0.0098
6	GDS2667	Embryonic R1 stem cell differentiation in vitro (MG-430B)	36 h	*vs.*	7 d	time	0.539	0.0098
7	GDS1414	Candoxin effect on glial cells: time course	24 h	*vs.*	12 h	time	0.521	0.011
8	GDS2666	Embryonic R1 stem cell differentiation in vitro (MG-430A)	48 h	*vs.*	7 d	time	0.518	0.011
9	GDS1823	Embryonic stem cell differentiation induced by various chemicals: time course	DMSO	*vs.*	RA	agent	0.513	0.0119
10	GDS2668	Embryonic J1 stem cell differentiation in vitro (MG-430A)	18 h	*vs.*	7 d	time	0.512	0.012
11	GDS2666	Embryonic R1 stem cell differentiation in vitro (MG-430A)	36 h	*vs.*	7 d	time	0.509	0.0122
12	GDS1733	Heat shock transcription factor HSF1 depleted cells response to heat shock: time course	24 h	*vs.*	4 h	time	0.505	0.0127
13	GDS2666	Embryonic R1 stem cell differentiation in vitro (MG-430A)	48 h	*vs.*	9 d	time	0.505	0.0127
14	GDS1688	Various lung cancer cell lines	small cell cancer	*vs.*	non-small cell adenocarcinoma	cell line	0.491	0.0146
15	GDS2666	Embryonic R1 stem cell differentiation in vitro (MG-430A)	36 h	*vs.*	9 d	time	0.487	0.0152

Figure 5 Search results for Nanog knockdown.

Application to FoxO3 knockout in neural stem cells

To evaluate the predictive potential of our search method in another system, we examined the effects of FoxO3 knockout in neural stem cells (NSCs). FoxO3 regulates NSC homeostasis by preventing premature differentiation and controlling oxygen metabolism [40]. Throughout the body, the FoxO family of transcription factors regulate a wide variety of cellular processes including glucose metabolism, cell cycle arrest, differentiation, and detoxification of reactive oxygen species (ROS) [41,42]. We created a DE profile comparing wild type to *FoxO3*$^{-/-}$ adult mice using normalized data from GSE18326. A query of GEO DE profiles yielded numerous significant results, the most significant of which are shown in Figure 7. Several matching profiles (Results 2, 3, 5 and 9) implicate FoxO3 in hypoxia response: data from GDS2758 and GDS2760 compare MCF-7 breast cancer cells under hypoxic and normoxic conditions as well as with siRNAs targeting hypoxia-inducible factor 1 (HIF-1α) and HIF-2α. Bakker et al. found that FoxO3 is activated in response to hypoxic stress in mouse embryonic broblasts (MEFs), and furthermore that this activation requires functional HIF-1α [43]. Renault et al. also found that FoxO3 is required for the expression of hypoxia-dependent genes in NSCs [40]. GDS2162 (Result 5) compares p300 and CBP null MEFs in response to dipyridyl (DP) or control (EtOH). DP, a hypoxia mimetic, induces HIF-1α [44] and thus potentially FoxO3. In all four hypoxia-related profile matches, therefore, the direction of the comparisons accurately predicts known FoxO3 biology. To further probe the relationship between FoxO3 and hypoxia, we examined the genes responsible for the high correlation between our FoxO3 query and Result 2. Predictably, we found genes associated with both hypoxia and FoxO3 signaling (Figure 6A). For instance, DDIT4 and NDRG1, both of which have been found previously to be activated during hypoxia [45,46] also contain FoxO binding motifs in their regulatory regions [40].

Other matches from the FoxO3 search (Results 11, 12, 15) point to a role for FoxO3 in cellular response to cytokine interleukin-2 (IL-2) stimulation. All three of these matching profiles compare cytotoxic T cell

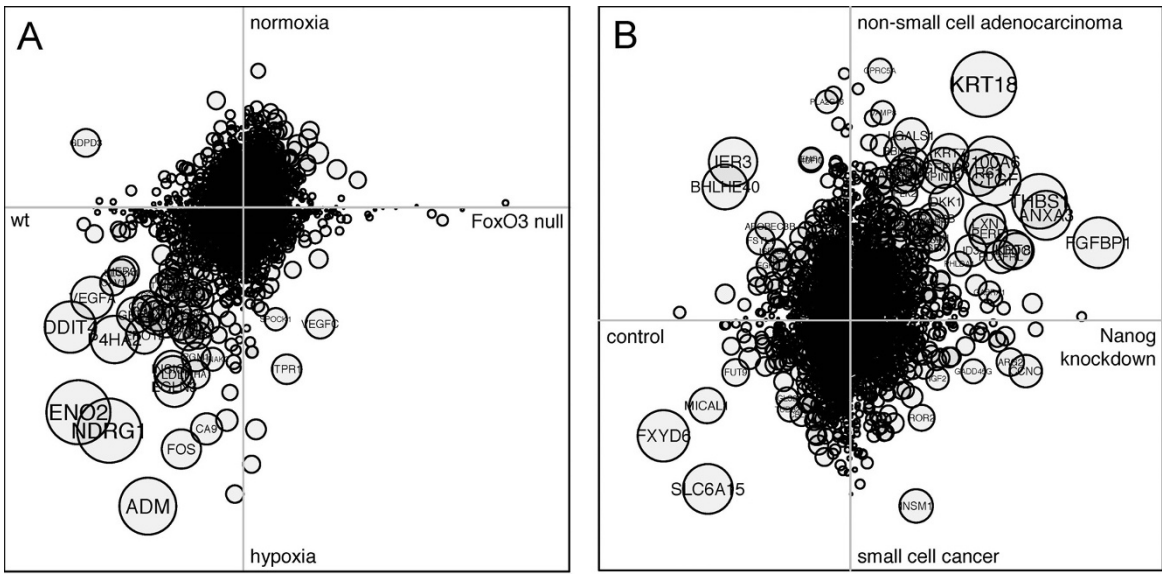

Figure 6 Gene-level analysis of transcription factor search results. Scatterplots comparing the log fold-change of each gene shared by two differential expression profiles. Expression values are centered and scaled. The area of each circle is proportional to the contribution that the gene makes to the final correlation score, and thus is a function of the magnitude as well as the significance of differential expression. (A) Comparison of GSE18326: FoxO3 null versus wild type and GDS2758: normoxia versus hypoxia. (B) Comparison of GDS1824: Nanog knockdown versus control and GDS1688: non-small cell adenocarcinoma versus small cell cancer.

line (CTLL-2) at 1 hour after IL-2 stimulation to a later time point (6, 12, or 16 hours). From the direction of these comparisons, we would predict that IL-2 stimulates a transcriptional program that is similar to that of FoxO3 knockout. Indeed, IL-2 signaling leads to phosphorylation and inactivation of FoxO3 in CTLL-2 cells [47], confirming this hypothesis.

Query: GSE18326 Role of FoxO3 in adult neural stem cell maintenance in mice — Wild type *vs.* FoxO3 null

Rank	GEO	Title	Subset 1	*vs.*	Subset 2	Type	Score	*q-value*
1	GDS2106	Lymphoblastoid cell lines from various CEPH pedigrees	CEPH pedigree 1444	*vs.*	CEPH pedigree 1345	genotype /variation	0.756	0.0013
2	GDS2758	Hypoxia and 2-oxoglutrate-dependent dioxygenase inhibition (HG-U133A)	hypoxia	*vs.*	normoxia	protocol	0.746	0.0015
3	GDS2760	Hypoxia-inducible factor depletion (HG-U133 2.0)	control	*vs.*	HIF-1alpha depletion	protocol	0.730	0.0017
4	GDS2106	Lymphoblastoid cell lines from various CEPH pedigrees	CEPH pedigree 1444	*vs.*	CEPH pedigree 1340	genotype /variation	0.723	0.0018
5	GDS2162	CH1 domain deletion, p300 and CBP heterozygous null mutant hypoxic fibroblasts response to trichostatin A	DP	*vs.*	EtOH	agent	0.722	0.0018
6	GDS998	Treacher Collins' syndrome Tcof1 gene overexpression and knockdown effect on neuroblastoma cells	N1E-115 wild type	*vs.*	N1E-115 Tcof1 overexpressed	cell line	0.712	0.0021
7	GDS2106	Lymphoblastoid cell lines from various CEPH pedigrees	CEPH pedigree 1447	*vs.*	CEPH pedigree 1345	genotype /variation	0.710	0.0022
8	GDS2106	Lymphoblastoid cell lines from various CEPH pedigrees	CEPH pedigree 1447	*vs.*	CEPH pedigree 1340	genotype /variation	0.700	0.0024
9	GDS2760	Hypoxia-inducible factor depletion (HG-U133 2.0)	HIF-2alpha depletion	*vs.*	HIF-1alpha depletion	protocol	0.698	0.0024
10	GDS587	Myogenic differentiation timecourse (MG-U74C)	4 d	*vs.*	-2 d	time	0.694	0.0025
11	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	1 h	*vs.*	12 h	time	0.693	0.0025
12	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	1 h	*vs.*	6 h	time	0.684	0.0026
13	GDS2106	Lymphoblastoid cell lines from various CEPH pedigrees	CEPH pedigree 1444	*vs.*	CEPH pedigree 1341	genotype /variation	0.681	0.0026
14	GDS2666	Embryonic R1 stem cell differentiation in vitro (MG-430A)	18 h	*vs.*	6 h	time	0.671	0.0029
15	GDS3222	Cytotoxic T cell line response to interleukin-2: time course	1 h	*vs.*	16 h	time	0.670	0.0029

Figure 7 Search results for FoxO3A knockout.

Discussion

In optimizing our data processing and search pipeline, we found that linear combinations of gene expression features derived in a separate compendium benefited our analysis. The most effective dimension reduction technique involved projecting each DE profile into a feature-space identified by independent component analysis. We previously used ICA to identify fundamental components of human gene expression from a large compendium of 10,000 arrays, of which only a small subset overlap with the experiments examined here [27]. The ICA projection method reduced the set of features from on the order of 20,000 to less than 500, allowing for rapid indexing and searching of large libraries of differential expression profiles. Furthermore, this approach outperformed module-based methods, possibly because the linear model incorporated data from all of the genes rather than only those that participate in discrete gene sets. Despite the fact that these ICA features were derived in human data, they proved robust in identifying and ranking experiments in closely related species as well. Thus, our results support previous findings that gene expression features derived in one compendium can be useful for interpreting data from new datasets [48].

To calculate similarities between differential expression profiles, we introduced a novel weighting scheme

that incorporates information about a feature's significance of differential expression. This approach provides an intuitive means for emphasizing the contributions of features that are significantly differentially expressed in both experiments, which may represent the most relevant common biology. At the same time, the weighted correlation incorporates even genes that are not significantly differentially expressed, potentially capturing the effects of broader transcriptional changes. We observed that this scheme worked well with Pearson correlation, but did not perform as well when combined with rank-based correlation. Future work will characterize the behavior of this similarity measure on a larger scale.

We used the most successful data processing pipeline to index all GEO DataSets. Our results with transcription factor experiments suggest that this approach can provide predictions for genes, phenotypes and perturbations that share functional similarities with a query experiment. Analysis of Nanog knockdown in ESCs successfully identified other ESC differentiation time courses, induced by a variety of factors, from amongst almost 10,000 other profiles (Figure 5). The same search predicted a link between small lung cell carcinoma and ESC transcriptional programs. For a less well characterized transcription factor, FoxO3, our method also succeeded in recapitulating known biology across species and experimental systems. Although it is clear that FoxO3 has lineage-specific effects [42,49], we identified a role for FoxO3 in hypoxia response that appears to transcend tissue type [40,41]. For uncharacterized comparisons, this information has the potential to provide useful hypotheses for phenotypes and pathways to investigate.

As in more traditional microarray analyses, however, interpretation of the most significant genes identified by our weighting scheme remains difficult. Our analysis of the FoxO3 search revealed a number of genes involved in both hypoxia and FoxO3 signaling, linking these two pathways. However, the top genes in the Nanog knockdown search failed to reveal convincing pathways that might explain the relationship between small lung carcinoma and ESCs. While we focus on the interpretation of several individual genes in this study, future efforts may benefit from the use of gene set enrichment tools to find pathways that are significantly represented in the top gene list.

As experimentalists continue to explore and deposit information about cellular processes and perturbations, the utility of content-based search approaches will increase. With a larger bank of transcriptomic data and a high chance of identifying overlapping and functionally related biology, an "experiment-omic" screen might be the first step in characterizing a novel dataset. To realize this, further ontological indexing of expression databases may also be necessary [50]. Several groups have already begun to integrate expression with textual phenotype data to enable gene function prediction [51] and automatic disease diagnosis [14] from large databases. Even for expression-driven methods, controlled annotations for experimental variables, tissue types, and culture systems would allow for more accurate assessments of functional relevance. Finally, ontological indexing of textual annotations will enable the creation of more sophisticated connectivity maps linking not just diseases and drugs, but also gene knockdowns, over-expression studies, and genotype comparisons. These ontology-informed studies may not only search public repositories based on gene expression, but also provide meta-analysis across phenotypic categories.

Conclusions

We have explored computational methods needed to search large repositories for relevant experiments based on differential expression, using an experiment as a query. While previous studies use hard thresholding to select gene sets of interest [4,11,13], we propose a data-driven approach that uses information from all shared genes to compare two experiments. Differential expression profiles containing scores for each gene or feature were generated and compared using correlation metrics, following the hypothesis that this direct and intuitive method would perform well across diverse datasets. In a collection of 32 experiments comparing normal to diseased tissue, we achieved an average AUC of 0.737 for retrieving experiments that measure the same disease. We further demonstrated the ability of our method to identify functionally relevant experiments from a large database of studies. Future work will include implementing the principles learned here into a web-based application. Public deployment of these methods will enable discoveries in drug repurposing, disease classification, and systems repositioning as we explore the molecular underpinnings of diverse biological processes and phenotypes.

Methods

Disease compendium

From a previous collection of disease-associated NCBI GEO microarray experiments [5], we collected 1,278 processed arrays comprising 32 experiments that compared normal to diseased tissue for Duchenne muscular dystrophy, breast cancer, and Huntington's disease. These experiments represented a variety of species, platforms, tissues, and normalization techniques, factors which might strongly influence the clustering of expression data.

Differential expression profiles

In transcriptomic studies, differential expression analysis identifies the genes and biological processes that vary

between two samples. To represent this information from different datasets in a standardized manner, we mapped probesets to Entrez Gene identifiers using AILUN [52] and generated differential expression (DE) profiles for each comparison using Bioconductor software [53]. Here, a DE profile consists of a list of features (e.g., genes) each with an associated score (e.g., fold change). For each comparison, we represented this differential expression score in three ways.

Log fold-change profile

We converted all microarray data to log values by examining the maximum and minimum values of the normalized probe-level data and applying $\log_2$ transformation as needed. We aggregated probes to genes using the fixed effects meta-estimate, calculating an average for each gene weighted by the variance of each probe [54]. We calculated the fold-change difference between normal and disease by averaging samples within each group.

P-value profile

Probes were aggregated as for the log fold-change method. For each gene in each experiment, we determined the probability that the gene was differentially expressed with an empirical Bayes moderated *t*-statistic implemented in the *limma* R package (version 2.16.5) [55]. We corrected for multiple hypothesis testing using the Benjamini-Hochberg method [56]. For DE profiles represented in terms of a reduced set of features (see below), we applied *limma* to assess the differential expression of that feature.

Rank profile

For each sample in each experiment, we ranked probes based on their raw expression score, then averaged all scores for a probe to create a single score for normal and disease sample groups. We mapped from probes to genes by finding the median of the subtractive difference between all pairwise combinations of probes for the same gene in normal and disease.

Dimension reduction

For all three DE profile representations, we mapped genes to their human homologs using NCBI Homologene, removing genes that did not have one-to-one homologs between species (Additional file 7). While removing species-specific genes may result in loss of important biological information, we hypothesized that comparing global, conserved patterns of gene expression between experiments would prove sufficient to predict functional associations (see Additional file 1 for data on the number of genes mapped for each dataset discussed in the manuscript). Next, we applied one of two methods of dimension reduction.

Projection onto independent components

We previously used independent component analysis (ICA) to identify fundamental features in human gene expression space by analyzing a compendium of 9,460 heterogeneous human microarrays run on the Affymetrix HG-U133 Plus 2.0 platform [27]. Briefly, we applied hierarchical clustering to our compendium to normalize the contributions of over- or under-represented conditions, applied independent component analysis to the normalized data, and aggregated the results over 20 runs using the partitioning around medoids clustering algorithm [57]. The resulting 423 components provide a data-driven feature space on which to map new gene expression data. For each DE profile, we considered the common genes between the experiment and the gene-to-component mapping, then projected the DE profile into ICA feature space using:

$$A = S^T X, \tag{1}$$

where A is the final reduced profile (423 features), S is the component matrix (components × genes), and X is the original profile in gene-space.

Fixed effect meta-estimate

To evaluate the performance of the ICA projection method, we also used a set of known features to reduce the dimensionality of our DE profiles. Given a collection of gene sets, we calculated a meta-score for each gene set using the fixed-effect meta-estimate, which represents an average across all genes in the set weighted by their inverse variance [58]. This method summarizes the contributions of functionally coherent gene sets, and may be appropriate for expression analysis. We used MSigDB v2.5, a well described collection of 5,452 gene sets most often used in conjunction with GSEA [16]. For comparison, we also derived gene sets from the ICA features described above: for each independent component, we created a module from all genes that scored three standard deviations above the mean in one direction. These 423 modules represented data-derived functionally coherent gene sets as determined by GO enrichment [27].

Similarity measures

We compared DE profiles in gene- or feature-space using similarity measures based on Pearson correlation coefficient and Spearman rank correlation coefficient. We used three weighting schemes. First, unweighted correlations are typically used in microarray clustering and search applications. However, this approach does not incorporate information about the variability of a gene, either across the compendium or within each dataset. Thus we tested a weighting scheme that accounts for the magnitude as well as the variance of a gene's change. We reasoned that genes with high variability across the compendium should be weighted lower than genes with low variability; that is, a change of the same magnitude should be more significant for a gene with low variance than for a gene with high

variance, since its relative deviation from the mean would be higher. To account for this, we calculated an inverse-variance weighted correlation, where each feature is weighted by the inverse of its variance across the entire compendium. Finally, we explored the possibility of weighting each gene by a function of its differential expression in the two datasets. Intuitively, a gene that is differentially expressed in both datasets should receive more weight than a gene that is differentially expressed in one or neither dataset. While Pearson correlation already rewards high magnitude changes, we chose to further weight genes by their p-value of differential expression to incorporate the inter-dataset variance as well as the magnitude. We calculated the weights for this p-value weighted correlation using:

$$w_i = \left[-log\left(p_{i1} p_{i2} \right) \right]^{1/C}, \quad (2)$$

where w_i is the weight for feature i, p_{ij} is the FDR-corrected empirical Bayes p-value for experiment j, and C is a scaling factor. For this work, we empirically chose C = 2 because it delivered the best clustering of our disease compendium (data not shown). We used the ROCR package [59] to evaluate the performance of various data processing methods.

GEO DataSet search

To search GEO for experiments with similar transcriptional patterns, we indexed all GEO DataSets (GDSs). We downloaded processed data from GEO and used the GDS "Value type" field to transform the data to $\log_2$ space. Each GDS is manually annotated with one or more factors, e.g., "disease state" or "time", which outline the experimental conditions that vary between groups of samples. Within each GDS, we compared all combinations of groups for a single factor. For each of these comparisons, we created two DE profiles: one in gene-space, and one in the ICA feature-space described above. We calculated p-values for each gene and ICA feature using the empirical Bayes modified t-test as described [27]. To search these DE profiles, we used the absolute value of the p-weighted Pearson correlation metric, since the direction of the comparison is arbitrary. To assess the significance of DE profile comparisons, we selected 10,000 random pairs of comparisons to serve as a background distribution of correlation scores. We estimated the false discovery rate (FDR) of our search results by calculating the percentage of these random comparisons that exceed a given similarity score. Because this random sampling may include true positive comparisons (e.g., two profiles from the same dataset), our corrected p-values may underestimate the significance of new comparisons.

Additional material

Additional file 1: Evaluation of data representation methods. We explored three alternative methods for representing differential expression data: log fold-change, normalized rank difference, and adjusted p-value significance. Using our disease compendium, we performed leave-one-out cross-validation by using each of 32 experiments to query the others. We generated ROC curves with different combinations of data representation and correlation metrics. Bars and AUC estimates indicate standard errors for curves averaged over all cross-validation trials.

Additional file 2: Precision at 4 for ICA p-weighted Pearson search. We calculated the "precision at 4" for each experiment in the disease compendium. Red bars show null distribution creating by permuting labels with 95% confidence intervals.

Additional file 3: High resolution network of GEO differential expression profiles. Vector graphic representation of the network in Figure 4. See Additional file 4 for legend.

Additional file 4: Legend for differential expression profile network. Legend for Additional file 3.

Additional file 5: FXYD6 Expression in GDS1824. GEO Gene profile for FXYD6 in GDS1824. See http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo&term=GDS1824[ACCN]+fxyd6.

Additional file 6: FXYD6 Expression in GDS1688. GEO Gene profile for FXYD6 in GDS1688. See http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo&term=GDS1688[ACCN]+fxyd6.

Additional file 7: Homolog mapping. Excel Spreadsheet describing the number of genes mapped through Homologene to human for each dataset discussed in the text.

Acknowledgements

The authors thank Boris Oskotsky and Alex Skrenchuk for assistance and technical support; Nicholas Tatonetti for critical comments; and Anne Brunet and Ashley Webb for interpretation of FoxO3 results. Computing resources at the Stanford Center for Biomedical Informatics Research were funded by the Hewlett Packard Foundation and the Lucile Packard Foundation for Children's Health. Financial support was provided by the National Library of Medicine (R01 LM009719) and Howard Hughes Medical Institute.

Author details

[1]Department of Bioengineering, Stanford University School of Medicine, Stanford, CA, USA. [2]Biomedical Informatics Training Program, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Pediatrics and Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Computer Science, Stanford University, Stanford, CA, USA. [5]Departments of Genetics and Medicine, Stanford University School of Medicine, Stanford, CA, USA. [6]Lucile Packard Children's Hospital, Stanford University, Stanford, CA, USA.

Authors' contributions

JE, AB and RA conceived of the study. JE and AM designed the experiments. JD provided curated disease datasets. JE performed the experiments and wrote the manuscript. RT conducted preliminary experiments and implemented a prototype web-site. RC contributed code for generating GEO comparisons. All authors reviewed and approved the final manuscript.

Received: 30 June 2010 Accepted: 21 December 2010
Published: 21 December 2010

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286(5439)**:531-7.

2. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**:109-26.
3. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, Lerner J, Brunet JP, Subramanian A, Ross KN, Reich M, Hieronymus H, Wei G, Armstrong SA, Haggarty SJ, Clemons PA, Wei R, Carr SA, Lander ES, Golub TR: **The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease.** *Science* 2006, **313(5795)**:1929-35.
4. Hassane DC, Guzman ML, Corbett C, Li X, Abboud R, Young F, Liesveld JL, Carroll M, Jordan CT: **Discovery of agents that eradicate leukemia stem cells using an in silico screen of public gene expression data.** *Blood* 2008, **111(12)**:5654-62.
5. Dudley JT, Tibshirani R, Deshpande T, Butte AJ: **Disease signatures are robust across tissues and experiments.** *Mol Syst Biol* 2009, **5**:307.
6. Suthram S, Dudley JT, Chiang AP, Chen R, Hastie TJ, Butte AJ: **Network-based elucidation of human disease similarities reveals common functional modules enriched for pluripotent drug targets.** *PLoS Comput Biol* 2010, **6(2)**:e1000662.
7. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Edgar R: **NCBI GEO: archive for high-throughput functional genomic data.** *Nucleic Acids Res* 2009, , **37 Database:** D885-90.
8. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca-Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A: **ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression.** *Nucleic Acids Res* 2009, , **37 Database:** D868-72.
9. Hubble J, Demeter J, Jin H, Mao M, Nitzberg M, Reddy TBK, Wymore F, Zachariah ZK, Sherlock G, Ball CA: **Implementation of GenePattern within the Stanford Microarray Database.** *Nucleic Acids Res* 2009, , **37 Database:** D898-901.
10. Fujibuchi W, Kiseleva L, Taniguchi T, Harada H, Horton P: **CellMontage: similar expression profile search server.** *Bioinformatics* 2007, **23(22)**:3103-4.
11. Chen R, Mallelwar R, Thosar A, Venkatasubrahmanyam S, Butte AJ: **GeneChaser: identifying all biological and clinical conditions in which genes of interest are differentially expressed.** *BMC Bioinformatics* 2008, **9**:548.
12. Caldas J, Gehlenborg N, Faisal A, Brazma A, Kaski S: **Probabilistic retrieval and visualization of biologically relevant microarray experiments.** *Bioinformatics* 2009, **25(12)**:i145-53.
13. Kapushesky M, Emam I, Holloway E, Kurnosov P, Zorin A, Malone J, Rustici G, Williams E, Parkinson H, Brazma A: **Gene expression atlas at the European bioinformatics institute.** *Nucleic Acids Res* 2010, , **38 Database:** D690-8.
14. Huang H, Liu CC, Zhou XJ: **Bayesian approach to transforming public gene expression repositories into disease diagnosis databases.** *Proc Natl Acad Sci USA* 2010, **107(15)**:6823-8.
15. Hibbs MA, Hess DC, Myers CL, Huttenhower C, Li K, Troyanskaya OG: **Exploring the functional landscape of gene expression: directed search of large microarray compendia.** *Bioinformatics* 2007, **23(20)**:2692-9.
16. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci USA* 2005, **102(43)**:15545-50.
17. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus.** *Nat Rev Genet* 2006, **7**:55-65.
18. Raychaudhuri S, Sutphin PD, Chang JT, Altman RB: **Basic microarray analysis: grouping and feature reduction.** *Trends Biotechnol* 2001, **19(5)**:189-93.
19. Clarke R, Ressom HW, Wang A, Xuan J, Liu MC, Gehan EA, Wang Y: **The properties of high-dimensional data spaces: implications for exploring gene and protein expression data.** *Nat Rev Cancer* 2008, **8**:37-49.
20. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97(18)**:10101-6.
21. Liebermeister W: **Linear modes of gene expression determined by independent component analysis.** *Bioinformatics* 2002, **18**:51-60.
22. Saeys Y, Inza I, Larrañaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23(19)**:2507-17.
23. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, Georgescu C, Romero R: **A systems biology approach for pathway level analysis.** *Genome Res* 2007, **17(10)**:1537-45.
24. Liu CC, Hu J, Kalakrishnan M, Huang H, Zhou XJ: **Integrative disease classification based on cross-platform microarray data.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:S25.
25. Yeung KY, Medvedovic M, Bumgarner RE: **Clustering gene-expression data with repeated measurements.** *Genome Biol* 2003, **4(5)**:R34.
26. Sjögren A, Kristiansson E, Rudemo M, Nerman O: **Weighted analysis of general microarray experiments.** *BMC Bioinformatics* 2007, **8**:387.
27. Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB: **Independent component analysis: Mining microarray data for fundamental human gene modules.** *J Biomed Inform* 2010, **43**:932-44.
28. Small CL, Shima JE, Uzumcu M, Skinner MK, Griswold MD: **Profiling gene expression during the differentiation and development of the murine embryonic gonad.** *Biol Reprod* 2005, **72(2)**:492-501.
29. Wilhelm D, Palmer S, Koopman P: **Sex determination and gonadal development in mammals.** *Physiol Rev* 2007, **87**:1-28.
30. Loh YH, Wu Q, Chew JL, Vega VB, Zhang W, Chen X, Bourque G, George J, Leong B, Liu J, Wong KY, Sung KW, Lee CWH, Zhao XD, Chiu KP, Lipovich L, Kuznetsov VA, Robson P, Stanton LW, Wei CL, Ruan Y, Lim B, Ng HH: **The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells.** *Nat Genet* 2006, **38(4)**:431-40.
31. Mitsui K, Tokuzawa Y, Itoh H, Segawa K, Murakami M, Takahashi K, Maruyama M, Maeda M, Yamanaka S: **The homeoprotein Nanog is required for maintenance of pluripotency in mouse epiblast and ES cells.** *Cell* 2003, **113(5)**:631-42.
32. Chambers I, Colby D, Robertson M, Nichols J, Lee S, Tweedie S, Smith A: **Functional expression cloning of Nanog, a pluripotency sustaining factor in embryonic stem cells.** *Cell* 2003, **113(5)**:643-55.
33. Hailesellasse Sene K, Porter CJ, Palidwor G, Perez-Iratxeta C, Muro EM, Campbell PA, Rudnicki MA, Andrade-Navarro MA: **Gene function in early mouse embryonic stem cell differentiation.** *BMC Genomics* 2007, **8**:85.
34. Williams RL, Hilton DJ, Pease S, Willson TA, Stewart CL, Gearing DP, Wagner EF, Metcalf D, Nicola NA, Gough NM: **Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells.** *Nature* 1988, **336(6200)**:684-7.
35. Hibi K, Takahashi T, Sekido Y, Ueda R, Hida T, Ariyoshi Y, Takagi H, Takahashi T: **Coexpression of the stem cell factor and the c-kit genes in small-cell lung cancer.** *Oncogene* 1991, **6(12)**:2291-6.
36. Watkins DN, Berman DM, Burkholder SG, Wang B, Beachy PA, Baylin SB: **Hedgehog signalling within airway epithelial progenitors and in small-cell lung cancer.** *Nature* 2003, **422(6929)**:313-7.
37. Young GD, Winokur TS, Cerfolio RJ, Van Tine BA, Chow LT, Okoh V, Garver Rl Jr: **differential expression and biodistribution of cytokeratin 18 and desmoplakins in non-small cell lung carcinoma subtypes.** *Lung Cancer* 2002, **36(2)**:133-41.
38. Cauffman G, De Rycke M, Sermon K, Liebaers I, Van de Velde H: **Markers that define stemness in ESC are unable to identify the totipotent cells in human preimplantation embryos.** *Hum Reprod* 2009, **24**:63-70.
39. Delprat B, Schaer D, Roy S, Wang J, Puel JL, Geering K: **FXYD6 is a novel regulator of Na, K-ATPase expressed in the inner ear.** *J Biol Chem* 2007, **282(10)**:7450-6.
40. Renault VM, Rafalski VA, Morgan AA, Salih DAM, Brett JO, Webb AE, Villeda SA, Thekkat PU, Guillerey C, Denko NC, Palmer TD, Butte AJ, Brunet A: **FoxO3 regulates neural stem cell homeostasis.** *Cell Stem Cell* 2009, **5(5)**:527-39.
41. Tothova Z, Gilliland DG: **FoxO transcription factors and stem cell homeostasis: insights from the hematopoietic system.** *Cell Stem Cell* 2007, **1(2)**:140-52.
42. Salih DAM, Brunet A: **FoxO transcription factors in the maintenance of cellular homeostasis during aging.** *Curr Opin Cell Biol* 2008, **20(2)**:126-36.

43. Bakker WJ, Harris IS, Mak TW: **FOXO3a is activated in response to hypoxic stress and inhibits HIF1-induced apoptosis via regulation of CITED2.** *Mol Cell* 2007, **28(6)**:941-53.
44. Kallio PJ, Wilson WJ, O'Brien S, Makino Y, Poellinger L: **Regulation of the hypoxia-inducible transcription factor 1alpha by the ubiquitin-proteasome pathway.** *J Biol Chem* 1999, **274(10)**:6519-25.
45. Jögi A, Vallon-Christersson J, Holmquist L, Axelson H, Borg A, Påhlman S: **Human neuroblastoma cells exposed to hypoxia: induction of genes associated with growth, survival, and aggressive behavior.** *Exp Cell Res* 2004, **295(2)**:469-87.
46. Shoshani T, Faerman A, Mett I, Zelin E, Tenne T, Gorodin S, Moshel Y, Elbaz S, Budanov A, Chajut A, Kalinski H, Kamer I, Rozen A, Mor O, Keshet E, Leshkowitz D, Einat P, Skaliter R, Feinstein E: **Identification of a novel hypoxia-inducible factor 1-responsive gene, RTP801, involved in apoptosis.** *Mol Cell Biol* 2002, **22(7)**:2283-93.
47. Stahl M, Dijkers PF, Kops GJPL, Lens SMA, Coffer PJ, Burgering BMT, Medema RH: **The forkhead transcription factor FoxO regulates transcription of p27Kip1 and Bim in response to IL-2.** *J Immunol* 2002, **168(10)**:5024-31.
48. Tamayo P, Scanfeld D, Ebert BL, Gillette MA, Roberts CWM, Mesirov JP: **Metagene projection for cross-platform, cross-species characterization of global transcriptional states.** *Proc Natl Acad Sci USA* 2007, **104(14)**:5959-64.
49. Paik JH, Kollipara R, Chu G, Ji H, Xiao Y, Ding Z, Miao L, Tothova Z, Horner JW, Carrasco DR, Jiang S, Gilliland DG, Chin L, Wong WH, Castrillon DH, DePinho RA: **FoxOs are lineage-restricted redundant tumor suppressors and regulate endothelial cell homeostasis.** *Cell* 2007, **128(2)**:309-23.
50. Shah NH, Jonquet C, Chiang AP, Butte AJ, Chen R, Musen MA: **Ontology-driven indexing of public datasets for translational bioinformatics.** *BMC Bioinformatics* 2009, **10(Suppl 2)**:S1.
51. Malone BM, Perkins AD, Bridges SM: **Integrating phenotype and gene expression data for predicting gene function.** *BMC Bioinformatics* 2009, **10(Suppl 11)**:S20.
52. Chen R, Li L, Butte AJ: **AILUN: reannotating gene expression data automatically.** *Nat Methods* 2007, **4(11)**:879.
53. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10)**: R80.
54. Stevens JR, Doerge RW: **Combining Affymetrix microarray results.** *BMC Bioinformatics* 2005, **6**:57.
55. Smyth GK: **Linear models and empirical bayes methods for assessing differential expression in microarray experiments.** *Stat Appl Genet Mol Biol* 2004, **3**:Article 3.
56. Benjamini Y, Hochberg Y: **Controlling the false discovery rate: a practical and powerful approach to multiple testing.** *J Roy Statist Soc* 1995, **B(57)**:289-300.
57. Kaufman L, Rousseeuw PJ: *Finding groups in data: an introduction to cluster analysis* Hoboken, N.J.: Wiley; 2005 [http://www.loc.gov/catdir/enhancements/fy0626/2005278659-b.html].
58. Hedges LV, Olkin I: *Statistical methods for meta-analysis* Orlando: Academic Press; 1985 [http://www.loc.gov/catdir/description/els032/84012469.html].
59. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R.** *Bioinformatics* 2005, **21(20)**:3940-1.

doi:10.1186/1471-2105-11-603
Cite this article as: Engreitz *et al.*: Content-based microarray search using differential expression profiles. *BMC Bioinformatics* 2010 **11**:603.